DOCUMENT RESUME

ED 042 804	TM 000 076
AUTHOR	Brink, Nicholas E.
TITLE	Characteristics of Rasch's Logistic Model.
PUB DATE	Mar 70
NOTE	15p.; Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970
EDRS PRICE	EDRS Price MF-$0.25 HC Not Available from EDRS.
DESCRIPTORS	Item Analysis, *Mathematical Models, *Rating Scales, *Statistical Analysis, Statistics, *Testing
IDENTIFIERS	Guttman Model, *Rasch Logistic Model

ABSTRACT

A basic description of the Rasch model and a brief review of the work previously done on the model is related. Simulated data was used to test goodness of fit to the Rasch model. It was found that data with near perfect fit to the Guttman model provides perfect fit to the Rasch model, though a perfect Guttman scale collapses. Random data also provides a good fit. The effects of varying the standard deviations on normally distributed total scores and the ranges on uniformly distributed scores and the effects of combining two sets of independent data are analyzed in some detail. [Not available in hard copy due to marginal legibility of original document.] (DG)

Characteristics of Rasch's Logistic Model

Nicholas E. Brink

University of California, Los Angeles

Much has been said about the characteristics of data that affect the various statistics used in evaluating tests. Such knowledge is of importance in making decisions as to which model and which statistic should be used in such evaluations.

Rasch Model

Work has just begun in describing the characteristics of data that produce good fit to the Rasch model. In her dissertation Panchapakesan (1969) has explored the effect of varying item discrimination, the presence of "bad" items, and the effect of guessing on the model. She found that inequality of item discrimination, lack of unidimensionality, and variation in guessing decreased fit to the model. The statistic used in evaluating fit to the Rasch model is the chi-square test for goodness of fit between observed data and expected values of that data. For a clear presentation of the Rasch model and this test of goodness of fit see Wright and Panchapakesan (1969).

In asking the question, "What do data with perfect fit to the Rasch model look like?", characteristics of both the classical and Guttman models were considered. Similarity was noted between the Guttman and Rasch models. While the Rasch model scales items on easiness and subjects on ability, the

Guttman model orders the items on difficulty and the subjects on total score.

This similarity does not continue, however. The Rasch model alone determines estimates of subject ability and item easiness that have ratio scale characteristics, i.e. the ability scale has a zero point that means zero ability and a zero point on the item easiness scale means no easiness or infinite difficulty. For the Rasch model the ability estimates are independent of the item easinesses and the number of items. This is not the case with Guttman scaling where scores are the total number of correct responses and thus dependent upon both the number and difficulty of items. The Rasch scaling procedure also produces ability estimates independent of the sample of subjects used to scale or calibrate the item easinesses. This independence has been well illustrated by Wright (1968). The Guttman model does not attempt to make assumptions more restrictive than that the data are ordinal on the two dimensions - i.e. item difficulty and total score. Conversely, Rasch's model is a latent trait model, a model that estimates the persons underlying trait while the Guttman model only produces a score relative to the items used.

One of the findings of Panchapakesan (1969) is that equal item discrimination is necessary for good fit to the Rasch model. Figure 1 represents the item characteristic curves (ICC) for 5 items over a range of ability. These curves represent the probability of responding correctly to an item if given the subject's ability. For equal item discrimination, these curves

need to be parallel. This is not necessary for Guttman scaling. Instead Guttman scaling requires that these curves be discrete or nonoverlapping on the range of ability measured by each item (Figure 2).

Simulation I: Perfect Guttman Scale

In that the Rasch model is a function of two independent factors, the model is relatively complex and the influences of various factors that may enter the model are difficult to determine. For this reason, at the 1969 AERA pre-session on Rasch scaling, Wright suggested using simulated data. The present study followed this suggestion. This also provided a way to control for unwanted variables and to provide a way to compare the two models. In submitting simulated perfect Guttman data to the Rasch analysis a problem developed. In that items either answered correctly or missed by all subjects are beyond the range of the calibration sample the easiness of these items cannot be determined. When this occurs the items are deleted. Similarly, when a subject misses all items or answers all items correctly the ability of these subjects exist somewhere beyond the range of ability measured by this set of items and thus cannot be determined. Again these subjects are deleted from the response matrix. With perfect Guttman data, either at least one item or at least one subject is deleted from the response matrix, leaving another subject or item to be deleted on the next round of truncation. The result

is no data to be analyzed by the Rasch model.

Simulation II: Near Perfect Guttman Data and Random Data

To avoid this problem, near perfect Guttman data were generated with enough random deviation from a perfect Guttman scale so that few or no items or subjects would be deleted. This was done by introducing a normally distributed random error variable into the model. The effect of increasing the standard deviation of this error was to increase the probability (from zero to one half) of a subject missing an item below his ability level and, similarly, to increase the probability of passing an item above his ability level. In that this error was normally distributed this probability diminished as the distance of the easiness of the items from an item measuring the subjects ability level increased. With the error standard deviation being small the data would appear to be near a perfect Guttman scale. In this case, the data produced a fit to the Rasch model with a probability of one.

The nature of the probabilistic model needs to be examined here. In that the Rasch model is a probabilistic model, distributions of probabilities of item and person parameters are assumed and produced. If there is no variability in the observed scores from the expected scores then the expected probabilities of the model are not met. Thus this probabilistic model is not appropriate for the data. This is what occurs when the chi-square probability of fit reaches one. This may be better seen with an example.

If six coins were repeatedly tossed you would not expect to get exactly three heads and three tails on each toss. If this were the case the probability distribution of these events would not be met.

In the case of a perfect Guttman scale no variability is allowed. This is seen in the definition of reproducibility. Reproducibility is the necessary condition for a perfect Guttman scale, where from a person's total score his response pattern to the set of items can be exactly determined. In this sense the Guttman model is a deterministic model contradicting the assumptions of the Rasch model. The saught after probability for the chi-square test of goodness of fit is one half. Great deviation either way from one half represents a lack of fit to the Rasch model.

Data sets were then generated with increases in error deviation from a perfect Guttman scale. Fit to the Rasch model rapidly decreased to around a probability of .5 (Table 1). To explore this further, 20 sets of completely random data were generated and subjected to the Rasch analysis. These data sets represented the responses of 1000 Ss to 64 items. The average probability of goodness of fit was .414 with little variation, S.D. = .178, not significantly different from a probability of fit of .500.

This finding was at first unexpected -- what do such data mean? Wright's Law school Admission Test study (Wright,

1968) used data that had zero fit to the model which seems to be the more common occurrence with real data. In examining the original assumptions of the Rasch model this persistance of good fit for random data can be explained.

Discussion

Rasch bases his model on three assumptions:

"1. To each situation of a subject (ν) having to solve a test item (i), there corresponds a probability of a correct answer which we will write in the form $p = \lambda_{\nu i}/(1+\lambda_{\nu i})$, $\lambda_{\nu i} > 0$,

2. The situational parameter $\lambda_{\nu i}$ is the product of two factors $\lambda_{\nu i} = \varepsilon_\nu \varepsilon_i$ (ε_ν pertaining to the subject, ε_i to the item)".

These parameters have been translated to Z and E respectively by Wright (1968).

"3. All answers, given the parameters, are stochastically independent." Rasch (1966).

This term stochastically independent can be translated to "local independence." Though neither Rasch nor Wright uses this term in his writing it was used at the 1969 AERA pre-session on Rasch scaling. The term local independence is also used to describe a basic assumption of Birnbaum's model (Panchapakesan, 1969; Lord and Novick 1968). Birnbaum's model is similar to Rasch's model but with the additional parameter of item discrimination. The assumption of local independence provides that "at a fixed point X

the probabilities for joint occurrence are products of the separate probabilities (Lazarsfeld, 1960, p. 85)." In other words, "those examined at a given ability level who answer a given item correctly are no more likely to answer other items correctly than are those examinees at the same ability level who answer the given item incorrectly (Lord, 1966, p. 25)." Larzarsfeld (1969, p. 497) also said that "if a class of people are alike in an underlying property then the indicators of this property should not be statistically related in this class."

This can explain why completely random data show good fit to the model. It represents the case of "local" data, i.e. of subjects with the same ability level and all items at the difficulty level measuring that ability. In this case the item characteristic curve for each item would coincide and thus not meet the cirterion of a perfect Guttman scale.

This illustrates another advantage of the Rasch model over the Guttman model. The ideal and most precise estimates of a person's ability are made when the easiness of the items match the subjects ability, i.e. with repeated measurement of the persons' ability. With the Guttman model a person's ability is only measured by a single item or within an interval of only two items. The precision of measurement is determined by how fine of a discrimination can be made between these two items. If the items exactly meet his ability level then you would expect a 50-50 chance of the person passing each item. In this case the perfectness of the Guttman scale would be lacking.

Simulation III: Distribution of Total Score

Another factor that can be varied in simulating data is the distribution of total scores of the subjects. In that Guttman scaling is an ordinal procedure this should have no effect on reproducibility. Similarly, in that the Rasch model makes no assumpitons about the nature of the distribution of total scores, variation in the form of the distribution should have no effect on goodness of fit. Normal distributions varying as to standard deviation and uniform distributions varying as to range were examined. As expected, no systematic differences in fit as well as no differences in values of the ability estimates were found. This last point again illustrates the independence of ability estimates from the item easiness parameters and sample of subjects used to calibrate the ability scale.

Simulation IV: Two Factor Data

The last characteristic to be examined in this paper is the effect of combining two sets of independent data on fit to the model. Three sets of data composed of 1000 subjects responding to 32 items were generated with moderate deviation from the perfect Guttman scale. The distribution of total scores was uniform for all three sets. For two sets, named A and B, the rnage of total scores was 32. Except for error fluctuation set C was generated with a range of zero and with a mean of 16. Both data sets A and B had a chi-square probability of fit equaling 1.00. Although both sets of data were generated from the same distribution of total scores the two scores assigned to each subject were random; thus the total scores representing

the two factors were independent. The results of combining these two data sets were surprising. When sets A and B were combined the probability for the chi-square test of goodness of fit remained 1.00. When data sets A and C were combined the resultant chi-square probability was zero. Data set C by itself had a fit of .868. This lack of a fit of probability of 1 in data set C was due to the fact that, with truncation, six items were eliminated from the data set thus producing a relatively greater amount of deviation from the perfect Guttman scale.

Discussion:

In an attempt to explain these results several sections of Wright and Panchapakesan's computer program for the Rasch analysis (1968) were examined. One of the first matrices examined was the score group by item matrix. Elements of this matrix represent the number of times each item was answered correctly for each group of subjects receiving the same total score. When ordered as to difficutly, items from each set alternated. In the case of combining sets A and B the effect was to produce two items of similar discrimination and easiness when separately there had only been one for each set. Thus the only change in the nature of the data was the addition of items to the set. The new distribution of score groups or total scores was no longer uniform but approached a normal distribution with a range of possible scores from 2 to 64. From the central limit theorm, if an infinite number of these samples has been added, rather than just two, the distribution of total scores would have been normal (Wilks, 1962). But in that

this model is independent of the distribution of total scores this would have had no effect. The effect of differences between total score for the two subsets of items for each subject was lost. In examining item discrimination each item was similar (Fig. 3); again not contradicting the Rasch model. The question still remains: "What does this characteristic of the Rasch model mean when two independent sets of items combined yield a perfect fit?"

The second combination of sets of items, i.e. sets A and C, produced a zero fit. Sets A and B were composed of items with approximately equal discrimination; the discrimination of the items of set C were poorer. This was seen from examining the slopes of the ICC for each item. Thus when sets A and C were combined the new set had items of varying discrimination and thereby contradicted the Rasch model (Figure 4).

As was illustrated in this last simulation the independence of total scores is alone not enough to reduce fit to the model, again the important factor was item discrimination.

Conclusion:

The Rasch model and its chi-square test for goodness of fit is highly sensitive to changes in item discrimination, although the slope of the ICC as a measure of item discrimination in itself has no effect on goodness of fit. Variations in the distribution of total scores has no effect on goodness

of fit. Finally in considering two factor data, the meaning of the dimensionality of a data set seems to mean something different from that of classical test theory. What it means needs further study and explanation.

Handout

Characteristics of Rasch's Logistic Model

Nicholas E. Brink

Item Characteristic Curves

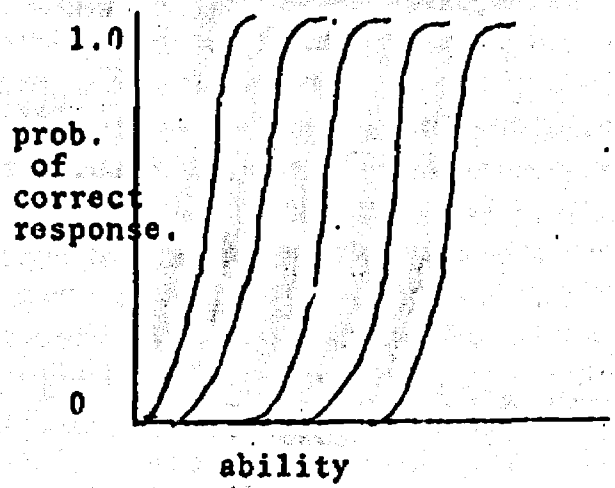

Figure 1

Perfect Rasch Scale

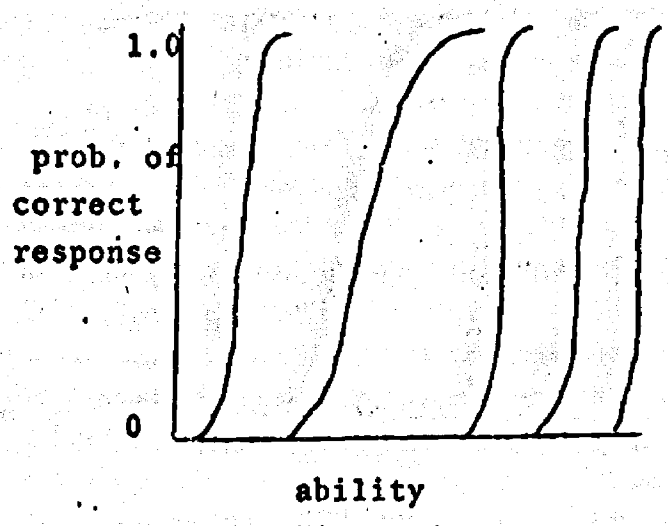

Figure 2

Perfect Guttman Scale

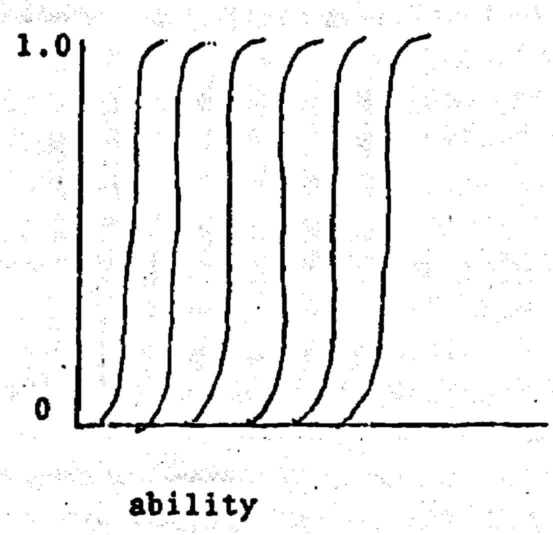

Figure 3

Data sets A and B combined.
(only 6 items illustrated)

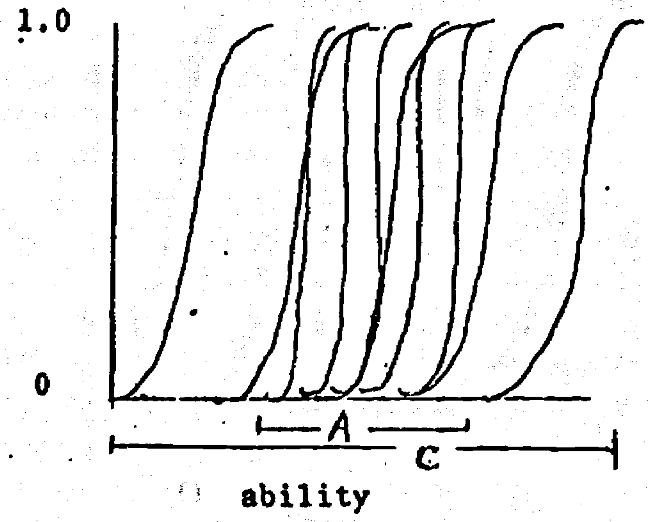

Figure 4

Data sets A and C combined.
(only 10 items illustrated.)

Table 1: Mean* Chi-square Probabilities for Eleven Different Degrees of Error Deviation from the Perfect Guttman Scale.

Error Deviation	Mean Chi-square Prob.
8	.986
16	.864
24	.691
32	.581
40	.628
48	.427
56	.428
64	.580
128	.582
160	.480
320	.565
Random Data**	.414

* Eight data sets were generated at each level of error deviation, these means differed at a .001 level of significance.

** This mean was based on 20 sets of Random Data.

References:

Guilford, J. P. Psychometric Method. New York: McGraw-Hill, 1954.

Lazarsfeld, P. F. Latent Structure Analysis. Psychology: A Study of a Science. Vol 3. Edited by Koch. New York: McGraw Hill, 1959 pp 276-543.

Lazarsfeld, P.F. Latent Structure Analysis and Test Tehory. Psychological Scaling: Theory and Applications. Edited by Gulliksen and Messick. New York: Wiley, 1960, 83-95.

Lord, F. M. Estimating Test Reliability. Educ. Psychol. Measurement, 15, 1955, pp. 325-336.

Lord, F. M. The Relation of Test Score to the Trait Underlying the Test. Readings in Mathematical Social Science. Edited by Lazarsfeld and Henry. Chicago: Science Research Assoc., 1966, 21-53.

Lord F. M. and M. R. Novick Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Mosier, C. I. On the Reliability of a Weighted Composite. Psychometrika, 1943, 8, 161-168.

Panchapakesan, N. The Simple Logistic Model and Mental Measurement. Dissertation, University of Chicago, 1969.

Rasch, G. An Individualistic Approach ot Item Analysis. Readings in Mathematical Social Science. Edited by Lazarsfeld and Henry. Chicago: Sceince Research Associates, 1966, 89-107.

Stouffer, S., et al. Measurement and Prediction. New York: Wiley, 1950.

Symond, P. M. Factors Influencing Test Reliability. Educational and Psychological Measurement. Edited by Payne and McMorris.

Waltham, Mass.: Blaisdell, 1967, 46-53.

White, B. W. and E. Saltz. The Measurement of Reproducibility. _Problems in Human Assessment_. Edited by Jackson and Messick. New York: McGraw-Hill, 1967, 241-257.

Wilks, S. _Mathematical Statistics_. New York: Wiley, 1963.

Wright, B. D. Sample-Free Test Calibration and Person Measurement. _Proceedings of the 1967 Invitational Conference on Testing Problems_. Princeton: Educational Testing Service, 1968, 85-101.

Wright, B. D. and N. Panchapakesan. _Unconditional Maximum Liklihood Item Analysis, Computer Program_. University of Chicago, 1968.

Wright, B. D. and N. Panchapakesan, Procedure for Sample-Free Item Analysis. _Educational and Psychological Measurement_, 1969, 29, 23-48.